SOUND EVENT LOCALIZATION AND DETECTION USING FOA DOMAIN SPATIAL AUGMENTATION

Technical Report

Luca Mazzon[1,2]*, Masahiro Yasuda*[1,*]*, Yuma Koizumi*[1] *and Noboru Harada*[1]

[1]NTT Media Intelligence Laboratories, Tokyo, Japan
[2]University of Padova, Padua, Italy

ABSTRACT

This technical report describes the system participating to the DCASE 2019, Task 3: Sound Event Localization and Detection challenge. The system consists of a convolutional recurrent neural network (CRNN) reinforced by a ResNet structure. A two-stage training strategy with label masking is adopted. The main advancement of the proposed method is a data augmentation method based on rotation in the first order Ambisonics (FOA) domain. The proposed spatial augmentation enables us to augment direction of arrival (DOA) labels without losing physical relationships between steering vectors and observations. Evaluation results on development dataset show that, even though the proposed method did not use any ensemble method in this experiment, (i) the proposed method outperformed a state-of-the-art system published before the submission deadline and (ii) the DOA error has significantly decreased: 2.73° better than the state-of-the-art system.

Index Terms— Sound event detection, direction of arrival estimation, CRNN, first order Ambisonics, data augmentation

1. INTRODUCTION

Sound event detection and localization (SELD) is the joint task of sound event detection (SED) and direction of arrival (DOA) estimation. SED task consists in recognizing the presence of certain sound classes in a potentially polyphonic audio recording, as well as their onset and offset times. DOA estimation consists in estimating azimuth and elevation angles of a sound source in an audio recording. The joint task of SELD, thus, requires to recognize, at each time frame, which sound classes are active and, for each of them, estimating the spatial coordinates of the corresponding sound source.

SELD is a challenge task of detection and classification of acoustic scenes and events (DCASE) 2019 Challenge, Task 3: Sound Event Localization and Detection [1]. The dataset and baseline system for the task are described in details in [2]. This baseline system was first introduced in [3] along an extensive study and comparison between the existing baseline methods and in different recording conditions. More recently, Cao et al. published a renewed SELD system [4] using a two-staged strategy, significantly improving the scores of the baseline.

Our system is based on Cao et al.'s system [4]. The differences from [4] are (i) network architecture, (ii) data augmentation, and (iii) ensemble methods, and these are describes in section 3.1, 3.2 and 3.3, respectively. The main advancement of the proposed method is a data augmentation method based on rotation in the first order Ambisonics (FOA) domain. The proposed spatial augmentation enables us to augment DOA labels without losing physical relationships between steering vectors and observations.

*The proposed system is the result of the conjunct work of L. Mazzon and M. Yasuda, who worked together and evenly cooperated on the task.

2. CONVENTIONAL METHOD

2.1. Problem setting

Let us define the SELD task. Here we define the STFT-spectrogram of the m-th microphone as $\boldsymbol{X}^{(m)} \in \mathbb{C}^{F\times T}$ and the set of M microphones' $\boldsymbol{X}^{(m)}$ as $\mathbf{X} = \{\boldsymbol{X}^{(m)}\}_{m=1}^{M}$, where T and F are the number of time-frames and frequency-bins, respectively. Given a number C of target events, the SELD task can be defined as the estimation problem of the c-th event's activity $z_{c,t} \in \{0, 1\}$, azimuth $\phi_{c,t} \in \mathbb{R}_{[-\pi,\pi)}$ and elevation $\theta_{c,t} \in \mathbb{R}_{[-\pi/2,\pi/2]}$ at time-frame t. Thus, the goal of the SELD task is designing a function for accurately estimating $z_{c,t}, \phi_{c,t}, \theta_{c,t}$ from $\mathbf{X}$.

2.2. The baseline system

SELDnet [2], the baseline system of the task, is a deep-neural-network (DNN)-based estimator of the target variables. The SELDnet uses a convolutional recurrent neural network (CRNN) which branches into two fully connected blocks, one with a sigmoid activation function for classification-based estimation of SED labels $z_{c,t}$, one with a linear activation function for regression-based estimation of DOA labels $\phi_{c,t}$ and $\theta_{c,t}$. That is, the SED part outputs a set of variables $r_{c,t}$ and then the presence probability of the c-th event at time frame t is estimated by using the sigmoid function as $p_{c,t} = p(z_{c,t}|\mathbf{X}) = \text{sigmoid}(r_{c,t})$. Finally, when $p_{c,t}$ exceeds the pre-defined threshold $0 \leq \alpha \leq 1$, the system identifies the c-th event as active at time frame t as

$$\hat{z}_{c,t} = \begin{cases} 1 & \text{for } p_{c,t} > \alpha \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

The acoustic features being used are magnitude and phase of the spectrogram of the four channels of either the FOA or microphone array dataset. Hereafter, we call these datasets FOA and MIC, respectively.

2.3. Cao et al.'s system

The new system introduced by Cao et al. [4] is still a SELDnet as a core structure but it introduced some key improvements. The first is

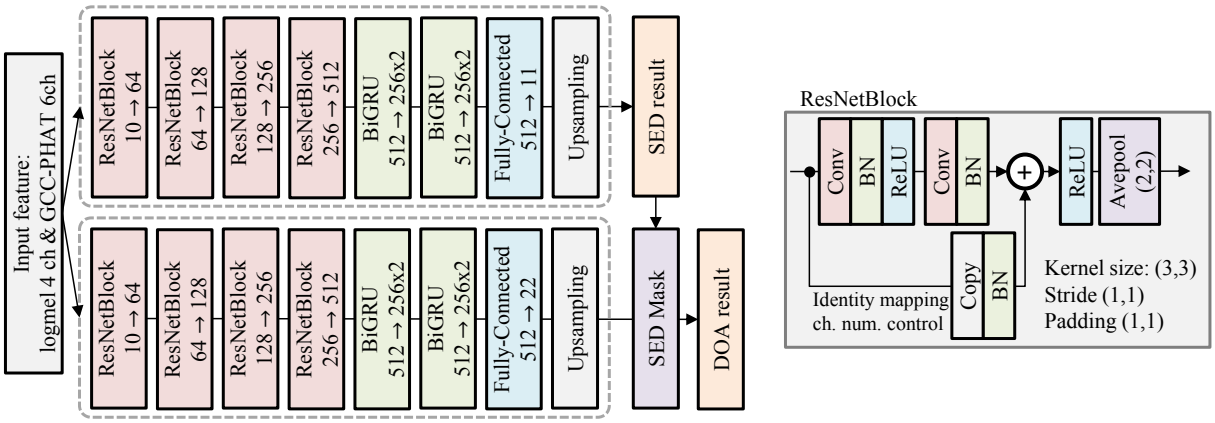

Figure 1: Proposed Network based on [4]. Left and right figure shows overall network architecture and detail of ResNetBlock, respectively. "Conv", "BN", and "Avepool" in right figure denotes convolutional layer, batch normalization, and average pooling, respectively.

on acoustic features. It uses the logmel magnitude spectrogram with $M = 96$ mel bins and the generalized cross correlation phase transform (GCC-PHAT) [5]. Since the logmel spectrum doesn't carry the phase information, which is important for DOA estimation, Cao et al. use the GCC-PHAT as a set of additional acoustic features.

Another key idea in [4] is using a two-staged training strategy, i.e. at first training only the SED branch of the network and in a second stage transferring the parameters of the CNN blocks, responsible for computing high level features, to the DOA branch and training it separately. During training of the DOA branch, SED ground truth labels are used to mask the estimated DOA labels. During inference, SED and DOA are estimated by the two separately trained branches and DOA estimated labels are masked by the estimated SED labels. This strategy has the goal of simplifying the training process while still keeping the advantages that SED features provide to the DOA estimation task. The CNN block architecture also differs from the baseline system, especially in the 2x2 pooling layer which shrinks the features along the time axis, with a subsequent up-sampling at the end.

3. PROPOSED METHOD

The proposed system uses [4] as a benchmark. In the following subsections, we'll describe the differences between [4] and our system.

3.1. Network architecture

Figure 1 (left) illustrates the overall architecture of our system. CNN blocks are capable of extracting high level features which are good both for SED and DOAE task. Thus, our first improvement has been adding an extra bi-directional gated recurrent unit (Bi-GRU) layer between the CNN blocks and the final fully connected layers, in order to allow the CNN blocks to keep this high level of feature computation and to reinforce the interpolation capability from these features. However, increasing the complexity of the model also implies two disadvantages: gradient vanishing and overfitting. To avoid the first problem, we employ a ResNet CNN structure [6] for each of the convolutional blocks, as shown in Fig. 1 (right). To address the second problem, we use a new data augmentation method which is described in the next section.

3.2. Data augmentation using FOA domain spatial augmentation

For improving the score we increased the domain representativeness of the dataset by using data augmentation. Data augmentation has been a widely used and successful strategy for most of DCASE challenge tasks [7–10] and other sound event detection tasks such as anomaly detection in sounds [11, 12]. However, to the best of the authors' knowledge, no augmentation strategy exists for DOA estimation. In the conjunct task of SED and DOA, there are some critical aspects to consider. First of all, when augmenting data, both SED and DOA information may be affected, thus respective labels must be updated correctly. For example, mixup augmentation [14] is a good strategy for conventional DCASE tasks. However, DOA labels in the regression format cannot be mixed up effectively[1]. Amplitude modulation and phase shifting applied differently on channels affects DOA information in a hardly predictable way. To overcome this problem, we propose a new augmentation strategy that allows us to increase the number of direction of arrivals represented in the dataset, as well as class-DOA combinations, while still being able to correctly compute the corresponding ground truth DOA labels of the augmented data. To do this, we exploited the simple equations describing the directional responses (steering vectors) of FOA channels.

We recall that, as described in the task description, the reference system is right handed with x axis pointing forward, y axis pointing leftwards and z axis pointing upwards, with azimuth angle ϕ increasing counterclockwise from x if seen from above and elevation angle θ increasing upwards from the horizontal plane xy. Given this coordinate system, for a given azimuth angle ϕ and a given elevation angle θ, the spatial frequency responses of the four FOA channels are the following:

$$\begin{aligned} H_1(\phi, \theta, f) &= 1, \\ H_2(\phi, \theta, f) &= \sqrt{3} * \sin\phi * \cos\theta, \\ H_3(\phi, \theta, f) &= \sqrt{3} * \sin\theta, \\ H_4(\phi, \theta, f) &= \sqrt{3} * \cos\phi * \cos\theta, \end{aligned} \tag{2}$$

where $*$ indicates multiplication. FOA channels correspond to the

[1] According to the original mixup strategy, labels are required to be in one-hot vector encoding [14]

Table 1: Sixteen patterns of simple spatial augmentation. X, Y, Z corresponds to channel H_3, H_4, H_2, respectively.

	$\phi_{c,t} - \pi/2$	$\phi_{c,t}$	$\phi_{c,t} + \pi/2$	$\phi_{c,t} + \pi$
$\theta_{c,t}$	$X \leftarrow Y, Y \leftarrow -X$	original	$X \leftarrow -Y, Y \leftarrow X$	$X \leftarrow -X, Y \leftarrow -Y$
$-\theta_{c,t}$	$X \leftarrow Y, Y \leftarrow -X, Z \leftarrow -Z$	$Z \leftarrow -Z$	$X \leftarrow -Y, Y \leftarrow X, Z \leftarrow -Z$	$X \leftarrow -X, Y \leftarrow -Y, Z \leftarrow -Z$
	$-\phi_{c,t} - \pi/2$	$-\phi_{c,t}$	$-\phi_{c,t} + \pi/2$	$-\phi_{c,t} + \pi$
$\theta_{c,t}$	$X \leftarrow Y, Y \leftarrow -X$	$Y \leftarrow -Y$	$X \leftarrow Y, Y \leftarrow X$	$X \leftarrow -X$
$-\theta_{c,t}$	$X \leftarrow -Y, Y \leftarrow -X, Z \leftarrow -Z$	$Y \leftarrow -Y, Z \leftarrow -Z$	$X \leftarrow Y, Y \leftarrow X, Z \leftarrow -Z$	$X \leftarrow -X, Z \leftarrow -Z$

all-pass filtered source sound (W), the front to back difference (X), the left to right difference (Y) and the up to down difference. We note that H_1 is the directional response corresponding to channel W of FOA, H_2 to channel Y, H_3 to channel Z and H_4 to channel X, as they can be seen as the *projections* of the sound source on the Cartesian axes. Given this nature of FOA, we consider that data augmentation can be achieved by using a rotation matrix, like suggested also in [13]. However, with a general transformation, there is the possibility of augmented labels going out of the domain of elevation angles defined for this task, i.e. $[-40°, 40°]$. In order not to go out of range, we use only reflections for augmenting elevation, while for azimuth we use all the combinations of ϕ, $-\phi$ and rotations of $+90°$, -90° and 180°. In total, for each DOA, we obtain 16 combination of DOAs, that is the original one plus 15 new ones. All patterns are listed in Table 1. Augmented positions are illustrated in Fig. 2. These are the most straightforward transformations to compute, thus, for a simple implementation, we used only these ones. For instance, a rotation of the azimuth angle of $+90°$ and a reflection on the xy plane, are described by the following changes of variable:

$$\begin{cases} \phi' = \phi + \frac{\pi}{2} \\ \theta' = -\theta \end{cases}, \tag{3}$$

which can than be substituted in (2) to obtain an expression of the new spatial responses as functions of the original ones:

$$\begin{aligned} H_1'(\phi, \theta, f) &= H_1(\phi, \theta, f) \\ H_2'(\phi, \theta, f) &= H_4(\phi, \theta, f) \\ H_3'(\phi, \theta, f) &= -H_3(\phi, \theta, f) \\ H_4'(\phi, \theta, f) &= -H_2(\phi, \theta, f) \end{aligned}. \tag{4}$$

In our system, which is based on Cao et al.'s system, which loads the entire dataset on the initialization of the data generator, we computed the augmented dataset in time domain and extracted the features offline for each of the augmented waveforms. For memory limitations, all transformations are applied offline and then the data generator chooses randomly between one of them at each iteration. However, a better solution for memory management and for being able to use more augmented data would be to compute the augmentation online directly on the features, which is not always feasible either for computational costs or for the complexity of the features (e.g. GCC-PHAT).

3.3. Model ensemble

As an ensemble method, we used linear regression approach. Model ensemble was processed for SED and DOA independently. Hereafter, we define N as the number of models for ensemble.

For SED ensemble, DNN outputs of each class were mixed before taking the sigmoid activation as

$$p_{c,t} = \text{sigmoid}\left(\sum_{n=1}^{N} w_{c,n}^{\text{sed}} r_{c,t,n}\right), \tag{5}$$

where $w_{c,n}^{\text{sed}}$ is the regression coefficient for c-th class and n-th SELD model, and $r_{c,t,n}$ is $r_{c,t}$ of n-th SELD model. In our submission, $w_{c,n}^{\text{sed}}$ was trained to minimize the binary-cross-entropy.

For DOA ensemble, the output of azimuth and elevation of all classes were calculated simultaneously using a large regression matrix $\boldsymbol{W}^{\text{doa}} \in \mathbb{R}^{2C \times 2CN}$. Here we define $\boldsymbol{d}_{t,n} = (\boldsymbol{\phi}_{t,n}^{\top}, \boldsymbol{\theta}_{t,n}^{\top})^{\top}$ as a vector which denotes a set of estimated azimuth and elevation of all classes by n-th SELD model. Then, the ensemble output of azimuth and elevation is calculated as

$$\boldsymbol{d}_t = \boldsymbol{W}^{\text{doa}} \left(\boldsymbol{d}_{t,1}^{\top}, ..., \boldsymbol{d}_{t,N}^{\top}\right)^{\top}. \tag{6}$$

In our submission, $\boldsymbol{W}^{\text{doa}}$ was trained to minimize masked mean-absolute-error (MAE) used in Cao et al.'s system as

$$\mathcal{L}^{\text{doa}} = \frac{1}{\mathcal{Z}} \sum_{t=1}^{T} \sum_{c=1}^{C} z_{c,t} \left(|\phi_{c,t} - \hat{\phi}_{c,t}| + |\theta_{c,t} - \hat{\theta}_{c,t}|\right), \tag{7}$$

where $\mathcal{Z} = \sum_{t=1}^{T} \sum_{c=1}^{C} z_{c,t}$.

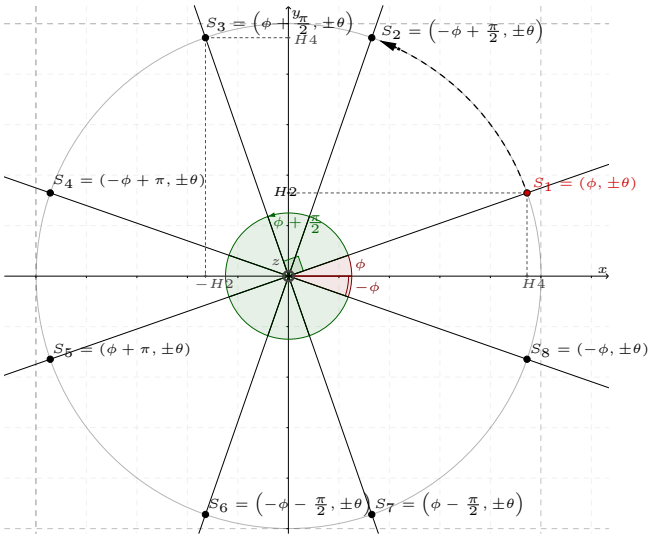

Figure 2: Augmented positions of the source S. Coordinates ϕ and θ are illustrated from an eagle eye view of the 3D space. For each of them, there is one with elevation coordinate $+\theta$ and one with elevation coordinate $-\theta$. Azimuth coordinate ϕ and its negative $-\phi$ are translated by $\frac{\pi}{2}$, π and $-\pi$.

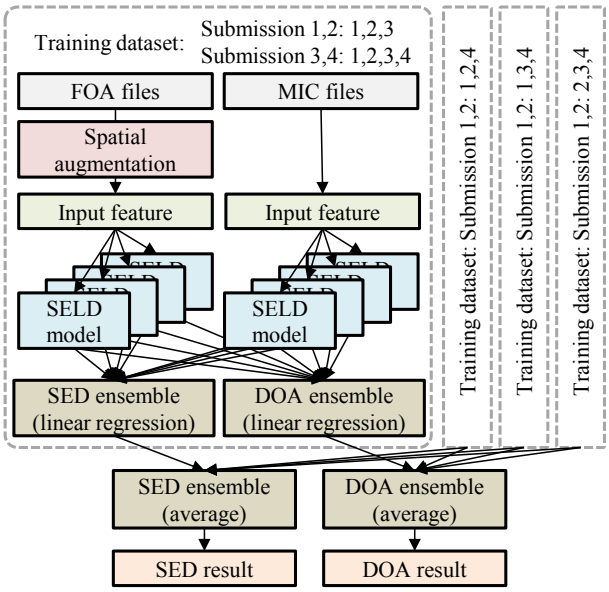

Figure 3: System overview.

3.4. Variations of submitted system

Figure 3 shows 4 variations of our submission. In submission 1 and 2, $N = 8$ SELD models with different initial parameters were trained using each of the 4 cross-validation patterns of the development dataset. 4 of the 8 SELD models used MIC wav files, and the others use FOA wav files which were augmented by FOA domain spatial augmentation. Early stopping was adopted for SELD model training using validation dataset, which was not used in SELD model training. Then, linear regression-based ensembles were adopted on the outputs of the 8 DNNs for each cross-validation pattern, so that we obtained 4 SED/DOA outputs each corresponding to a cross-validation pattern. Note that in submission 1 the regression parameters were trained using the three training splits, while in submission 2 the regression parameters were trained using the validation split. Finally, the final SED/DOA outputs were calculated as the average over the 4 cross-validation patterns.

In submission 3, $N = 8$ SELD models with different initial parameters were trained using all the wav files in the development dataset. In the same manner of submission 1 and 2, 4 SELD models used MIC wav files and the others used FOA wav files, which were augmented by FOA domain spatial augmentation. Then, a regression-based ensemble was used for both SED and DOA outputs. Since this model was trained using all the wave files in the development dataset, the final average-based ensemble used in submission 1 and 2 was not used. SED and DOA networks were trained for 40 and 50 epochs, respectively.

In submission 4, only one SELD model was trained using all the wav files in the development dataset. The input acoustic features were calculated from FOA wav files which were augmented by FOA domain spatial augmentation. In submission 4, we haven't used any ensemble method, and SED and DOA networks were trained for 40 and 50 epochs, respectively.

Table 2: Evaluation results on developement dataset. "ER", "F", "DOA", "FR", and "SELD" and means error rate, F-score [16], DOA error, Frame recall [17], and SELD score, respectively.

Name (split)	ER	F	DOA	FR	SELD
Baseline (all)	0.350	0.800	30.8°	0.840	0.220
Cao (all)	0.167	**0.909**	9.85°	0.863	0.112
Ours (all)	**0.166**	0.907	**7.12°**	**0.864**	**0.109**
Ours (1)	0.143	0.918	7.05°	0.871	0.097
Ours (2)	0.166	0.911	6.98°	0.863	0.108
Ours (3)	0.146	0.919	7.20°	0.872	0.099
Ours (4)	0.209	0.880	7.24°	0.849	0.130

3.5. Hyperparameters

In all submissions, the sample rate of STFT is set to 32kHz. A 1024-point Hanning window with a hop size of 320 points is utilized. The number of mel-band filters and the number of delayed samples of GCC-PHAT is set to $M = 96$. The audio clips are segmented to have a fixed length of 2 seconds with a 1-second overlap for training. The Adam method [15] is used as optimizer, and the learning rate is set to 0.001 for the first 30 epochs and is then decayed by 10% every epoch.

4. EXPERIMENTS

We evaluated the proposed method on the development dataset. To fairly evaluate the accuracy on the development dataset, we trained 4 SELD models using the same training setting as [4], that is we used 300 FOA wav files to train each SELD model and haven't used early stopping or ensemble methods. In this evaluation, FOA was selected as input, and only one SELD model of SED and DOA networks were trained with 40 and 50 epochs, respectively.

Table 2 shows the evaluation results of the proposed method. In Table 2, "Ours" denotes the proposed method, "Baseline" denotes the baseline system published by the task organizers [2] and "Cao" denotes the system which is the benchmark system of the proposed method [4], respectively. As we can see in the results, in terms of SELD score, the proposed method outperformed conventional methods published before the submission deadline. Notably, the DOA error was significantly decreased: 2.73° better than Cao et al.'s system. We believe it is mainly due to the FOA domain spatial augmentation. As future work, we will try to confirm the effectiveness of FOA domain spatial augmentation using several neural-network architectures and more general formulations.

5. CONCLUSIONS

In this technical report, we described the system participating to the DCASE challenge 2019 task 3. Our system is based on Cao et al.'s system [4]. The differences from [4] were (i) network architecture, (ii) data augmentation, and (iii) ensemble methods. The main advancement of the proposed method is a data augmentation method based on rotation in the FOA domain. Evaluation results on development dataset showed that even though the proposed method did not use any ensemble method in this experiment, (i) the proposed method outperformed a state-of-the-art system published before the submission deadline, and (ii) the DOA error was significantly decreased: 2.73° better than the state-of-the-art system.

6. REFERENCES

[1] "Dcase2019 task 3: Sound event localization and detection. task description.," 2019.

[2] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Submitted to Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE)*, 2019.

[3] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, 2018.

[4] Y. Cao, Q. Kong, T. Iqbal, F. An, W.Wenwu, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," *arXiv preprint, arXiv: 1905.00268v2*, 2019.

[5] C. H. Knapp and G. Carter, "The generalized correlation method for estimation of time delay, 1976. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp.320–327, 1976.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] Y. Sakashita and M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," in *Tech. report of Detection and Classification of Acoustic Scenes and Events 2018 (DCASE) Challange*, 2018.

[8] I.-Y. Jeong and H. Lim, "Audio tagging system for DCASE 2018: Focusing on label noise, data augmentation and its efficient learning," in *Tech. report of Detection and Classification of Acoustic Scenes and Events 2018 (DCASE) Challange*, 2018.

[9] M. Lasseck, " Acoustic bird detection with deep convolutional neural networks," in *Tech. report of Detection and Classification of Acoustic Scenes and Events 2018 (DCASE) Challange*, 2018.

[10] T. Inoue, P. Vinayavekhin, S. Wang, D. Wood2, N. Greco, and R. Tachibana, "Domestic activities classification based on CNN using shuffling and mixing data augmentation," in *Tech. report of Detection and Classification of Acoustic Scenes and Events 2018 (DCASE) Challange*, 2018.

[11] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised Detection of Anomalous Sound based on Deep Learning and the Neyman-Pearson Lemma," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol.27-1, pp.212-224, 2019.

[12] Y. Koizumi, S. Murata, N. Harada, S. Saito, H. Uematsu, "SNIPER: Few-shot Learning for Anomaly Detection to Minimize False-Negative Rate with Ensured True-Positive Rate," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, 2019.

[13] M. Kronlachner and F. Zotter, "Spatial transformations for the enhancement of Ambisonic recordings," in *Proc. of the International Conference on Spatial Audio*, 2014.

[14] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv:1710.09412*, 2017.

[15] D. Kingma and J. Ba, "Adam: A Method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[16] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, 6(6):162, 2016.

[17] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. of 2018 26th European Signal Processing Conference (EUSIPCO)*, pp.1462–1466. 2018.